RADBOUD UNIVERSITY

MSc THESIS

Graph Representations of News Articles for Background Linking

Author:
P.J. BOERS

Supervisor:
Prof. dr. ir. A.P. DE VRIES

Second Reader:
Dr. ir. F. HASIBI

A thesis submitted in fulfillment of the requirements for the degree of Master of science in Computing Science

Institute of Computing and Information Sciences
Faculty of Science

September 17, 2020

Abstract

Context defines the circumstances in which things happen and is crucial to fully understand complicated events. This applies especially to news articles, in which an adjustment in context can change the perspective of a story dramatically. To better understand an article, news sites can provide a list of background recommendations. To advance in the development of these systems and to investigate how news is represented online, the Text and REtrieval Conference started a track dedicated to information retrieval and news. This work investigates the performance of a graph-based retrieval method for the task of background linking. That is, the retrieval of context enriching background articles for a specific topic article. In order to do so, we explored the representation of news articles in graph form, looked at the application of named entities in the news domain and examined relevance measures to increase ranking performance. The proposed models were compared to a state-of-the-art "bag-of-words" baseline, but no significant improvements in effectiveness were found.

Contents

Chapter 1

Introduction

"For a newcomer to a contentious issue, context-free news is like walking into a cinema halfway through the movie and trying to make sense of a complicated plot."

— Pesach Benson

Context defines the circumstances in which things are happening, and provides the perspective through which we interpret events. It is said that in order to fully understand something context is crucial, that is why journalists are taught to provide a frame of reference for each story they write [1]. Failing to do so might lead to misinterpretation of their narrative. This may result in conflict, unfounded anger or general miscommunication.

1.1 Online News

In the last few years, online news consumption has become more mainstream; a study by Pew Research showed that in 2016, 38% of American adults consumed their news online [2]. Findings from their study in 2018 even show that 68% of Americans are getting at least some of their news online [3]. With the increase in consumption of online news also comes an increase in online publications, shifting the attention from providers to stories. Given that anybody can create professional looking news stories, comprehending the context behind such stories becomes more difficult [4].

In the first place, reading news online has many advantages over reading a newspaper. Think of the instant delivery time, enhanced portability (smartphones, tablets, laptops), push notifications for important events, personalized recommendations et cetera. Unfortunately, the wide adaptation of online news also brings some new problems to light. Online news platforms often benefit from the ambiguity of shocking headlines or the misinterpretation of facts in order to attract more readers to their site. This leads to a poor understanding of the actual news.

Another problem is ignorance about the knowledge of a reader. Different groups of people need different background information, making it harder for journalists to provide the right general pieces of background material. This is amplified by the guidelines for article length that many news sites pursue, leaving an author with only a few lines to provide context. Needless to say that this is often not enough to discuss all the necessary background information.

The retrieval of relevant background articles may compensate for the lack of context in online news stories. The extra information can contribute to an increase in understanding and allows for a more sophisticated interpretation of events. Ultimately, this also contributes to a better idea of the trustworthiness of a news article.

1.2 TREC News

In order to explore methods for the retrieval of relevant background articles, the Text REtrieval Conference (TREC) started a track dedicated to the role of information retrieval in the news domain. The News track started in 2018 and has focused on: the ranking of entities in news articles, the linking of background articles, and more recently the wikification of specific text parts. This work focuses on the task of background linking, in which the goal is to obtain a ranking of relevant background documents for a set of 50-60 topic articles. In the long term this will contribute to the creation of a dataset with annotated news articles that can be used as evaluation data. Eventually, this will help to improve the overall understanding of background linking in news articles.

1.2.1 Background Linking

The recommendation of background articles differs from general news recommendation, as it does not aim to retrieve news articles that fit the interest of individual users. Such systems rather contribute to a reduction of contextual information than an expansion, because readers are often kept inside their so called "bubble". Background linking also differs from the linking practice some authors do themselves, in which they connect parts of their story to other self-written articles. Even though some of these articles may indeed provide adequate background material, the focus of the News track lies on the large-scale retrieval of background documents based on their content, using automatized retrieval methods.

1.2.2 Graph Approach

Text is unstructured and allows for a wide range of structured thoughts. In other words, there are many ways in which you can convey the same message or idea. This makes it hard to automatize the extraction of meaning from a text. Almost all previous submissions for the task of background linking were based on a "bag-of-words" approach, in which articles were retrieved based on the overlap between article content and a set of query terms. Unfortunately, there are some aspects in which this approach falls short, namely, the lack of dependencies between words and the undue focus on content overlap. Graph representations provide a solution for these matters by including connections between terms and allowing for alternative retrieval strategies. Additionally, graphs provide a mechanism to expose relations between events and actors in news articles, possibly explaining incentives behind certain behaviour.

These potential advantages over current methods and the lack of complete graph-based retrieval techniques in previous editions of the track have inspired us to investigate the potential of graphs in the context of background linking. This aligns with the interest of the Radboud University Information Retrieval group (RUIR) in graph-based approaches to information retrieval. This work focuses on the representation of news articles in graph form, the application of graph-based ranking algorithms, and the relevance measure for news graphs in the task of background linking.

This thesis considers the following research questions:

***RQ 1**: What is the performance of a graph-based retrieval model in the task of background linking?*

***RQ 1a**: What is the best graph representation of news articles for this task?*

***RQ 1b**: How can we find the most important part of a graph that represents a news story?*

***RQ 2**: Does the inclusion of a score for new information affect background linking effectiveness?*

***RQ 3**: Can we produce a diverse list of recommendations without decreasing effectiveness?*

Chapter 2

Related Work

This chapter gives a brief overview of the methods currently used for background linking and describes some of the best attempts as presented in previous TREC News editions. Additionally, the usage of graph representations in the field of information retrieval is examined as well as the role of named entities in journalistic context.

2.1 TREC News Editions

The majority of work that was submitted to the news track in previous years transformed the topic article into an ad-hoc search query. This query formed a description of the information need and was used to retrieve documents that matched the information need best [5]. Such a retrieval problem traditionally exists of a query (description of topic article) and a relevance criterion, which allows for the judgement on whether a document meets the information need or not.

Preceding submissions mostly used relevance models to evaluate the usefulness of retrieved background documents. Each document in the collection was compared with an earlier constructed query, and the documents that contained the best combination of query terms obtained the highest score. This score was based on the weights of the terms in both the query and the document. Therefore, relevance was based on how similar a document in the collection was to the query article.

Usually, queries consist of a relatively small number of terms that concisely represent the information need; the average query length in online search engines lies between two and three terms [6]. However, in this case we are dealing with entire news articles which serve as a potential query. Usually, retrieval effectiveness increases with query length, but using complete documents as a query has proven to be unfeasible [7]. Mainly so, because many retrieval tools do not allow queries to extend a specific number of terms, forcing participants to come up with a smaller representation of news articles.

In practice the majority of prior methods reduced the query document to a set of keywords that represented the idea behind the news article. A crucial question then is, how to select those keywords that give a good representation of a news article? There is room for creativity here, and different techniques have been proposed.

One of the simplest extraction technique used, is the selection of the first N words in an article. This can be backed up by best-practice protocols for beginning journalists, which state that most important information should be mentioned in the first

paragraph of an article [1]. This technique was indeed tested in 2018 by the Anserini team [8]. They extracted the first 1000 terms of an article to form a query and used it to retrieve a set of background articles. Even though this approach obtained reasonably good results, one cannot deny that a news article might contain pertinent information in later paragraphs. Effectiveness was increased by selecting the 1000 terms with the highest *tf-idf* score. This score denotes a term's frequency (tf) multiplied with the inverse of the number of documents that term occurs in (idf) and plays a crucial role in many retrieval algorithms. The obtained score greatly relies on the 'eliteness' principle, which states that each term has some form of expressional power in its corresponding document [9]. This power boils down to a form of 'aboutness'; a term is elite in a document if the document is in some sense about the concept denoted by that term. The intuition behind *tf-idf* is that the more a term occurs in a document the more that document is about the topic denoted by that term. The *tf-idf* principle was also used by ICTNET and htw saar to construct their queries in the 2018 edition of the track [10, 11].

In the edition of 2019, Radboud University used the same *tf-idf* score to obtain the top 100 most representative terms, and expanded them with a RM3 expansion model [12]. Their query expansion rebuilt the initial query by incorporating terms from top results as retrieved with the primary *tf-idf* query. The main advantage of the expansion was an improvement in query reach, which led to an increase in recall, i.e. more of the targeted documents were found.

Besides the extraction of words based on their *tf-idf* score there are many other methods to determine a text's most important keywords. Some used the presence of real-world entities in news articles as most representative terms from a news article, yet others made use of graph analysis techniques [13, 14]. Qatar University, for example, achieved good retrieval effectiveness with the analysis of cores in a graph representation of news articles. These cores gave an indication of node influence and highlighted the importance of different terms [15].

2.2 Graph Representations

The "bag-of-words" representations of news articles come with some drawbacks that can be resolved by graphs. These graphs allow for a richer representation of articles and provide a convenient overview of coherence and general connection of ideas within a news story. Also, they allow for more sophisticated comparisons between topic and collection articles.

Graphs are usually defined as a set of vertices (nodes) and edges (connections): $G = (V, E)$, where $|V|$ represents the number of vertices and $|E|$ the number of edges. The edges denote the connection between separate nodes and can have a variety of characteristics. For example, the orientation of a relation (directed or undirected) and the strength of a relation (weighted or unweighted). Since nodes form the foundation of a graph, they determine a large part of the graph's success.

N-grams or sentences are frequently used for the representation of nodes, but larger pieces of text are also optional. Whether an edge exists or not depends on the criteria that are used to define relationships between nodes. The combination of connections and nodes make is possible to apply ranking functions within a graph. Ranking algorithms can be used to determine the importance of individual nodes and are often used to generate a hierarchy. Important nodes generally have more and stronger connections to other (important) nodes.

Previous TREC News submissions have used graphs for the extraction of keywords. Those methods used graphs to find the top N most important nodes and extracted the corresponding terms to form a query. The importance of a node could be determined via various strategies, but mostly depended on the number of connections a node had.

Submission from Qatar University based connections between nodes on the cooccurence of them in a text. They applied a sliding window to the text body and title and linked words that fell into the same window. The obtained graph was then used to create subgraphs that made the influence of the most important terms visible. Graphs were eventually composed into different cores, which were used for the extraction of keywords [15, 16].

The existence of connections between terms also allows for the recalculation of term importance based on these connections. The simplest example of such an algorithm is TextRank [17]. This is a modification of the well-known PageRank algorithm, which can be used to score web pages based on their hyperlinks [18]. Subsequently, TextRank can be used to score nodes in a graph based on their edges. Methods like this are mostly build on the belief that influential nodes are more important than less influential nodes. This might be best understood by the analogy with a voting system in which each node votes for the 'importance' of other nodes and each vote weight is based on the importance of that node. The more votes a node obtains the more important it is itself. The original TextRank model extracts keywords based on the importance of the nodes, but usage can also be restricted to node scoring only. A few notable extensions to the TextRank algorithm are: PositionRank, TopicRank, SingleRank and MultiRank [19–22]. Smith College experimented with these variations in their 2019 submission [23]. They used the algorithm to create keyphrases that were executed as a weighted query in a probabilistic relevance model.

Although previous submissions to the News track primarily used graphs for the extraction of keywords, graph-based document retrieval methods also exist in conventional information retrieval. Truong et al. created a biparte graph which connected documents and queries to terms and calculated similarity between the two based on the overlap in edges [24]. The edges had weights according to their *tf-idf*

score in the corresponding document. Their method outperformed the Okapi baseline model from the Lemur toolkit [1]. Zhang et al. represented both query and documents as graphs and compared them on the similarity between nodes and edges [25]. More specifically, they used the general maximum common subgraph (GMCS) as a measure to calculate graph similarity [26]. While their approach showed promising results when compared to a traditional vector space model, it was not efficient enough to be applied on an entire corpus. Therefore, it limited its efforts to a set of candidate documents, reducing the overall computational costs.

2.3 Named Entities and Journalism

News articles often comprise narratives about events concerning specific entities. These entities generally refer to specific objects that exist, such as persons, organizations, or locations. We speak of named entities when a reference is made to a real-world entity. For example, "Donald Trump" or "Barack Obama" are named entities but "president" is not.

Named entities occur regularly in news articles and have played a big role in previous TREC News editions (entity ranking). A likely explanation for the wide prevalence of entities is that many stories cover the interaction of real-world objects within real-world events. Moreover, news articles typically contain a particular combination of named entities that partake in a specific news story. Researchers from the City University of London have used named entities as a lead to find relevant background articles [13]. They hinted at the use of journalistic questions as a framework for an article's main story, claiming that the central idea behind an article can be summarized in a set of W questions: Who, What, Where, Why and When? These questions are part of the 5W1H method, which is used to obtain key aspects of a story in a structured way. Hamborg et al. [27], not participating in TREC News, have shown results in which they successfully answer these questions using state-of-the-art techniques from the domain of question and answering. Nonetheless, their model is rather computationally expensive and is not suited to be used on large collections.

[1] https://www.lemurproject.org/

Chapter 3

Data

3.1 Washington Post Collection

News articles were retrieved from TREC's Washington Post Corpus. The Washington Post is one of the biggest news papers in the United States and is located in Washington D.C. The most recent version of this corpus (v3) contains 671,947 news articles and blog posts from January 2012 through December 2019. However, as a way to evaluate our experiments we required the usage of the 2019 release (v2) which contained 608,180 news articles and blog posts from January 2012 through August 2017. The use of this release was necessary in order to preserve query relevance assessments that were obtained during the 2019 background linking track.

The 2019 edition of TREC News' background linking task contained 60 topic articles for which on average 260 articles were annotated per topic article. The rest of the articles were not annotated, and were assumed to be irrelevant. Topic articles varied in subject matter ranging from: "How amphetamine use may be affecting our waterways" to "How not to flip out when flipping a house".

Topics contained a topic identifier (num), document identifier (docid) and a url pointing to the relevant news page. Topics were provided in the following format:

Figure 3 Example of a topic

<top>
<num> Number: 826 **</num>**
<docid>96ab542e-6a07-11e6-ba32-5a4bf5aad4fa**</docid>**
<url>https://www.washingtonpost.com/sports/nationals/the-min..**</url>**
</top>

The document identifier referred to the id of the document in the corpus. Documents were stored in JSON format and contained the following tags: *id*, *article url*, *title*, *author*, *publication date*, *text contents*, *article type* and *source*. For the task of background linking all wire service articles were assessed as not relevant, nor were editorials or articles that described opinions. In practice this meant that sections with: "The Post's View", "Opinion" and "Letters to the Editor" in their *kicker* field (inside *text contents*) had to be filtered out.

3.2 Relevance

Graded relevance assessments were stored in a query-relevance file, with the degree of relevance grades varying between 0, 2, 4, 8 and 16. TREC mentions the following score-explanation combinations for the task of background linking [4]:

- 0: The linked document provides little or no useful background information.
- 2: The linked document provides some useful background or contextual information that would help the user understand the broader story context of the query article.
- 4: The document provides significantly useful background or contextual information that would help the user understand the broader story context of the query article.
- 8: The document provides essential useful background or contextual information that would help the user understand the broader story context of the query article.
- 16: The document MUST appear in the sidebar otherwise critical context is missing.

Figure 3.1 shows the distribution of relevance grades per topic. The majority of the retrieved articles in the 2019 edition of the track do not provide much background information, nor does every topic article have a highly relevant background article.

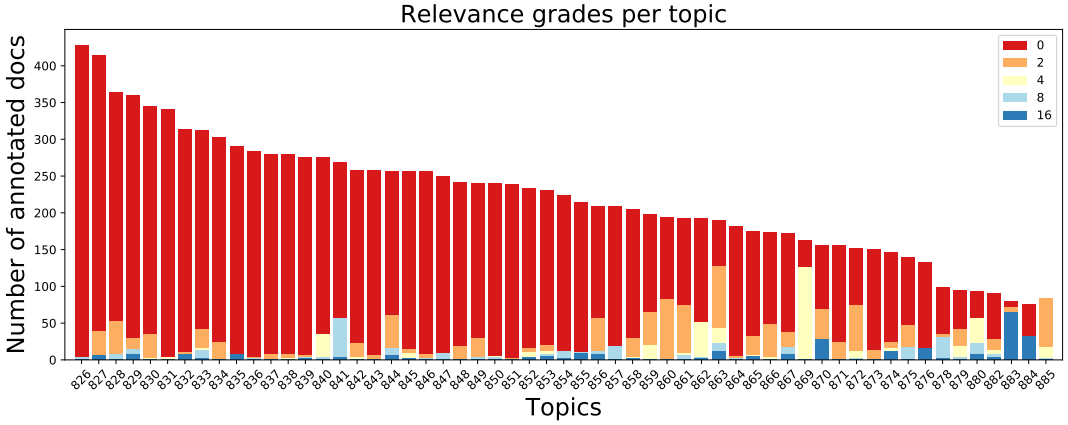

FIGURE 3.1: Distribution of relevance grades per topic article (2019)

Chapter 4

Methods

Graph-based retrieval methods have shown promising results in traditional retrieval tasks, and are able to outperform vector space models [25]. This work investigates the application of a graph-based retrieval model for the task of background linking. Deviation from previous attempts is obtained by replacing the widely used "bag-of-words" representation of news articles by a graph representation. The introduced connections between terms allow for new experiments in the task of background linking. Furthermore, the adoption of named entities in graphs and the role of entity types in some of the journalistic W-questions are investigated.

Trump Wants Greenland

To investigate the effectiveness of an entity driven graph-based retrieval system for the task of background linking, a model was developed for which different settings could be toggled on and off. This chapter describes the internal workings of this approach and explains the rationale behind chosen design principles. To obtain a clear view on the contribution of each choice, the following simplified article sample was used to visualize different procedures (fig 4).

Figure 4. Example snippet from a news article.

While he said that it was not a high priority, President Trump confirmed that he was looking into the possibility of acquiring Greenland which is an autonomous Danish territory.
Officials from Greenland and Denmark criticized the idea, and the Danish Prime Minister Mette Frederiksen eventually called the idea "absurd." She also tweeted that she "would have no interest in discussing the purchase of Greenland."[a]

[a] https://www.newsinlevels.com/products/trump-wants-greenland-level-3/

4.1 Baseline

The performances of the introduced graph methods were compared to Radboud's **BM25 + RM3** submission as presented in TREC 2019 [12]. Radboud's submission

was adopted as a baseline because it implemented a transparent and effective technique to retrieve background articles. Its simplicity and state-of-the-art performance made it an appropriate reference point for new methods. Their method is well suited to provide a set of candidate documents, as the architecture is highly optimized and scales appropriately. The reduction of computing time that was obtained by acquiring a set of candidate documents benefitted the proposed graph model by reducing the number of documents to process.

The baseline model made use of a generated "bag-of-words" query in combination with the well-known BM25 relevance model [9]. The top 100 terms with the highest *tf-idf* score were extracted from a topic article to form a search query. Before extraction took place, all terms in the article were stemmed and filtered for stop words. Additionally, words that contained dots or those that were shorter than three characters were discarded. The generated query was then used to search the collection for matching documents. The 10 documents with the highest rank were used for RM3 query expansion, creating a new query. Finally, a renewed search over the entire collection was done with the improved query and a ranking of roughly 100 documents was retrieved.

Figure 4.1. Bag-of-words representation of sample article (top 5 *TF-IDF* terms in bold)

['while', 'he', 'said', 'high', 'prioriti', 'presid', 'trump', 'confirm', 'he', 'look', 'possibl', 'acquir', '**greenland**', 'which', '**autonom**', '**danish**', 'territori', 'offici', 'from', '**greenland**', 'denmark', 'critic', 'idea', '**danish**', 'prime', 'minist', '**mett**', '**frederiksen**', 'eventu', 'call', 'idea', 'absurd', 'she', 'also', 'tweet', 'she', 'would', 'have', 'interest', 'discuss', 'purchas', '**greenland**']

Figure 4.1 shows an example of the stemming and stop word removal in the example article, the five words in bold possessed the highest *tf-idf* score. This resulted in the following initial query: "**greenland danish mett frederisken autonom**". After performing query expansion, the initial query was extended with additional terms. An article discussing the autonomy of Greenland in combination with its tradition of ice fishing could have expanded the initial query with terms like: "catfish" or "halibut".

4.2 Graph Representation

The introduction of a graph-based retrieval method for the task of background linking, asked for a suitable graph representation of news articles. Different graph configurations were tested and have been compared in terms of effectiveness.

4.2.1 Nodes

Nodes form the foundation of a graph representation and should be chosen carefully. Previous submissions to the News track have shown that the top N *tf-idf* terms were a good choice for query forming [8, 12]. Here, these terms were chosen as nodes in our graph.

Following the work by Zhang et al. [25] a variation of the *tf-idf* score was used to set the initial node weight. This was done as the formula generally gave a good indication of a term's importance; therefore, contributing to a solid foundation of the graph. In this work term weights were derived from the frequency of their occurrence in a document and were divided by the number of other terms i that also existed in that document. Subsequently, the term frequency was scaled by the inverse document frequency (idf) of the term in the collection. The term frequency tf for term t in document d is shown in the equation below. Similarly, the idf based on term occurrence in collection c and the scaled final weight W_{td} are shown as well.

$$tf_{td} = \begin{cases} \dfrac{1 + log(f_{td} - 1)}{\sum_{i=1}^{n} f_{id}} & \text{if } f_{td} > 1 \\ \dfrac{1}{\sum_{i=1}^{n} f_{id}} & \text{if } f_{td} = 1 \end{cases} \tag{4.1}$$

$$idf_{tc} = log(\frac{|c| - f_{tc} + 0.5}{f_{tc} + 0.5} + 1 \tag{4.2}$$

$$W_{td} = tf_{td} \cdot idf_{tc} \tag{4.3}$$

Based on the premise that the most important terms of a story are mentioned in the beginning of an article, node weights were updated by taking the text position of their term into consideration [1]. Previous editions of the News track showed that the importance of *entities* could be derived from their location in a text with high accuracy [12]. Weights in this work were updated by identifying in which paragraph a term was located, and then calculating the inverse of the paragraph's index (for simplicity index started at 1). That is, if the term occurred in the first paragraph the weight gained a score of 1. If the term occurred in a later paragraph the weight increased with a smaller amount, see 4.4. If a term occurred in multiple paragraphs, the paragraph with the lowest index was used.

$$W_{td} = tf_{td} \cdot idf_{tc} + \frac{1}{index(t, d)} \tag{4.4}$$

4.2.2 Edges

Two strategies to create connections between nodes were adopted, one using statistical features and one using semantic features. Both allowed for a different value of connection strength, and made different connections possible. Connections formed an indication of the likeliness that nodes belonged to the same part of a story. Strategies were based on connections between nodes depended on the distance between the paragraphs in which terms occurred and the proximity between term embeddings in vector space.

The first strategy drew on the assumption that (1) words within a paragraph describe the same subtopic, and (2) successive paragraphs introduce each other. This resulted in the creation of a connection between terms appearing in paragraph B and terms that appeared in either paragraph A, B or C. The weights were determined as follows:

$$W_{t_1,t_2} = \begin{cases} 1 & \text{if } t_1 \text{ and } t_2 \text{ are in the same paragraph} \\ \frac{1}{2} & \text{if } t_1 \text{ and } t_2 \text{ are in consecutive paragraphs} \end{cases} \tag{4.5}$$

The second strategy relied on word embeddings to estimate a generic relation between words. Vector representations of these words, consisting of latent features, were used to obtain a similarity score between words. Two words regularly occurring in the same context or grammatical structure led to a similar vector, and is an indication of compatibility. The distance between two vectors (4.6) was used to specify a relationship between terms, i.e. the closer terms occurred in vector space the stronger they were connected. Embeddings were adopted from Gerritse et al. who used Wikipedia to generate word embeddings that included named entities [28]. The inclusion of named entities was especially convenient for the connection of entity nodes, more about this in section 4.3.

$$W_{t_1,t_2} = \cos(\vec{t_1}, \vec{t_2}) \tag{4.6}$$

Performance was tested using both measures separately as well as a combination of the two. Different graph representations of news articles were created by varying node and edge weight formulas. Figure 4.1 shows one of the possible graph representation for our article sample (weights are omitted from the figure for simplicity).

4.3 Named Entities

Named entities were adopted as an additional option for nodes in the graph representation. This was supported by the idea that whenever articles possessed the same combination of named entities, the articles were likely dealing with the same (or at least closely related) story. News stories generally contain a more or less unique combination of named entities; therefore, using these entities as the foundation of a

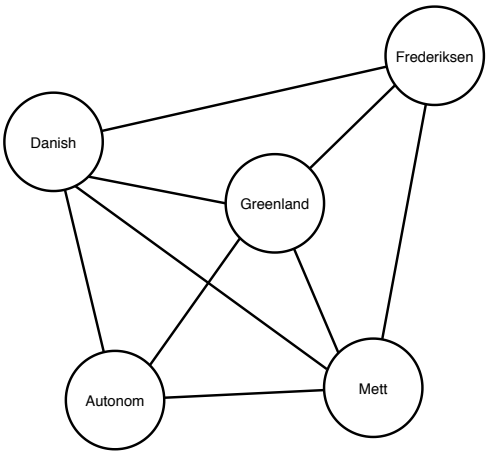

FIGURE 4.1: Graph representation of sample article.

graph seems adequate. Nonetheless, it is not unreasonable to assume that it is possible to make different stories with the same combination of named entities. Donald Trump, for example, executes many different activities in the same location within the same organization and among the same people. However, the mix of *tf-idf* terms and named entities might distinguish enough to differentiate between most articles.

Named entities were retrieved in a paragraph-wise manner using the REL entity detection module [29]. Entities were retrieved in such a way that all forms of one entity were mapped to the same base form. Additionally, entities were assigned a type that referred to their category of origin. Types varied between, PER (person), ORG (organization), LOC (location) or MISC (miscellaneous). The extraction of named entities in our sample article is shown in figure 4.3.

Figure 4.3. Named entities in example snippet.

While he said that it was not a high priority, President Trump (PER) confirmed that he was looking into the possibility of acquiring Greenland (LOC) which is an autonomous Danish (MISC) territory.
Officials from Greenland (LOC) and Denmark (LOC) criticized the idea, and the Danish (MISC) Prime Minister Mette Frederiksen (PER) eventually called the idea "absurd." She also tweeted that she "would have no interest in discussing the purchase of Greenland (LOC)."

4.4 TextRank

The methods described so far created graph representations based on weighted connections between terms/entities. However, there are techniques that can redefine a graph's most important nodes and edges based on an existing graph structure.

This work used the TextRank algorithm to update node weights based on their connections [17]. TextRank is typically used for the extraction of keywords in graphs,

but can also be used as a means to emphasize a graph's main story. The idea here is that by internally reranking the nodes (i.e. updating their weights), a more sophisticated estimate of an article's core story can be obtained. If two graphs contain the same node (e.g. "election") and both assigned it a high weight, then it is likely that the term fulfills a similar role in both articles. Finding overlap between multiple terms would then give an even better indication of story similarity.

The updated weight of a node was based on the weight of the incoming nodes multiplied with the weight of their connecting edges, divided by the number of outgoing edges of the incoming nodes (Eq 4.7). This was an iterative process that stopped after reaching a maximum number of iterations or when weight changes were no longer exceeding a predefined threshold. The default setting used a damping coefficient of 0.85, a convergence distance of 10^{-5} and 1000 iteration steps.

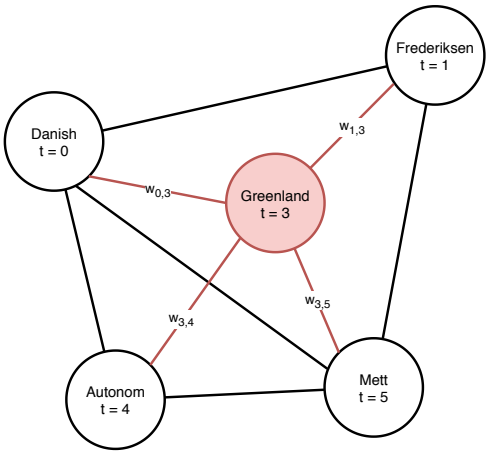

FIGURE 4.2: Illustration of TextRank input for "Greenland" node.

$$W_{t_i} = \sum_{j=1}^{n} \frac{W_{t_i,t_j}}{|edges(t_j)|} \cdot W_{t_j} \tag{4.7}$$

Before starting the iterations, all node weights were normalized in such a way that the sum of all nodes was equal to 1. We used n for the number of nodes j that had a connection with node i that was stronger than 0, i.e. $W_{t_i,t_j} > 0$. Effectiveness on the background linking task was tested both with and without the TextRank algorithm.

4.5 Graph Comparison

A combination of relevance measures was used to retrieve background documents that were both similar and dissimilar to the query article. That is, the retrieved documents should discuss the same topic, but should also include new information. The first relevance measure considered the overlap between graph nodes and edges, whereas the second focused on the detection of new story parts.

4.5.1 Greatest Maximum Common Subgraph

The similarity between query graph (Q) and candidate graph (C) was calculated in two steps. First, the weight of overlapping nodes was determined. That is, the weights of the nodes that occurred in both query and candidate graph (common subgraph, denoted as G_{CS}) were summed and divided by the sum of the nodes from the graph with the highest node weight score. Thereafter, the same procedure was performed for the overlapping edges, see equation 4.8. Eventually, the two scores were scaled with hyper parameter λ (in this work $\lambda = 0.5$).

$$sim(G_Q, G_C) = \lambda \frac{\sum\limits_{n_i \in G_{CS}} w_{n_i}}{max\left(\sum\limits_{n_i \in G_Q} w_{n_i}, \sum\limits_{n_i \in G_C} w_{n_i}\right)} + (1-\lambda) \frac{\sum\limits_{e_i \in G_{CS}} w_{e_i}}{max\left(\sum\limits_{e_i \in G_Q} w_{e_i}, \sum\limits_{e_i \in G_C} w_{e_i}\right)} \tag{4.8}$$

4.5.2 Novelty

As stated earlier, retrieving background articles based on similarity alone might not be sufficient to provide a reader with enough context. Identical articles normally do not expand one's understanding of a topic. As a means to ensure the detection of new information, a novelty measure was introduced as addition to the similarity measure.

In order to avoid the addition of irrelevant information, only those nodes that were strongly connected to the common subgraph (i.e. connection strength is above average) were considered to be a relevant contribution of novelty. On the contrary, if a node was part of the candidate article but was only weakly related to the subgraph, it was marked as irrelevant background information.

A background article was only found relevant if it entailed both a similar topic (similarity) and provided enough new information (novelty). Novelty was calculated by adding the weights of all novel nodes and dividing those by the sum of the nodes in the candidate article. To end up with a single score, we took the harmonic mean of the similarity and novelty scores; based on the hypothesis that if news articles were extremely similar, but did not deviate enough to provide new insights, or vice versa, it should obtain a low relevance score.

Figure 4.3 shows an example of a node that is linked to the subgraph of a candidate and sample article. The candidate article could have had information about the connection of Greenland's parliament with the Danish government. In this case the word "parliament" was strongly connected with the common part of the graphs and thus qualified as novel background contribution.

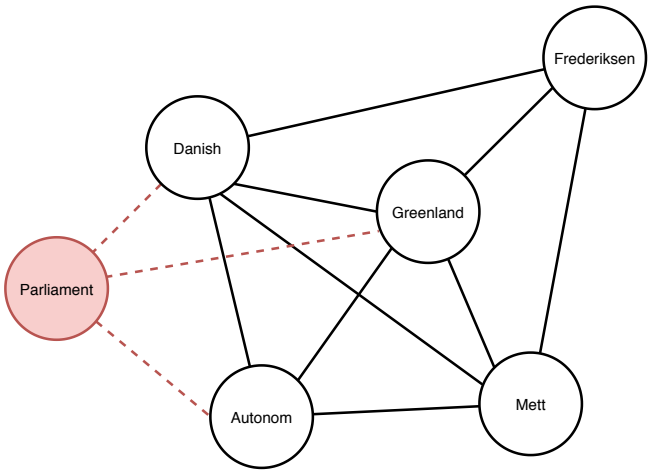

FIGURE 4.3: Example of novelty node.

4.6 Evaluation

4.6.1 Normalized Discounted Cumulative Gain

The background linking performance was measured using a normalized version of the discounted cumulative gain (NDCG), see equation 4.11. This is a measure of ranking quality and incorporates the assumption that relevant documents are worth more at higher ranks. A distinction between different amounts of relevance was used, as can be seen in chapter 3. Relevance value 0 did not contribute to the gain.

The DCG of a ranking with n documents could be obtained using formula 4.9. In order to normalize the score, the obtained DCG was divided by the ideal DCG. This is what you get if you have the perfect ranking of documents, and is shown in equation 4.10. $RANK_n$ is the set of documents in optimal order according to their relevance assessment.

$$DCG_n = \sum_{i=1}^{n} \frac{rel_i}{log_2(i+1)} \tag{4.9}$$

$$IDCG_n = \sum_{i=1}^{|RANK_n|} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{4.10}$$

$$NDCG_n = \frac{DCG_n}{IDCG_n} \tag{4.11}$$

4.6.2 Diversification

Although not actively evaluated, TREC prefers a diverse list of background documents over a monotone one. Even though there are no clear guidelines to define diversity, this work shows an attempt to bring some variety into the recommendations. News articles contain different entities from different categories (types). Assumed was that a diverse ranking of background articles consists of a set of articles with different entity types. Therefore, a new ranking was created in which each

document focused on a specific perspective of the story as denoted by the prevalence of its most dominant entity types (PER, ORG, LOC or MISC). The first three entity types aligned closely with the three most important journalistic W-questions: Who, What and Where? This provided us with journalistic insights about the angle of a story. Having an article with many *person* types probably covers an event from a perspective in which people play an important role (Who?), whereas an article that possesses mostly *location* types is considered with the whereabouts of a story (Where?). The support for a specific W-question that came with the retrieval of a particular article could then be identified by the prevalence of a specific entity type. Ultimately, this can help authors with directions for future work by identifying which perspectives are under represented.

A filter was used to discard documents if they provided the same kind of information as documents that were already in the ranking. For example, if the first document contained many entities with type *person*, consecutive documents were only stored in the ranking if they would focus on *location* or *organisation*.

4.7 Overview

A complete overview of the model can be found in figure 4.4. Note that for our experiments different configurations of the pipeline were tested. That is, separate runs were done for graphs that used named entities and those that did not. This also holds for the use of different node weight factors, graph similarity designs and filtering options.

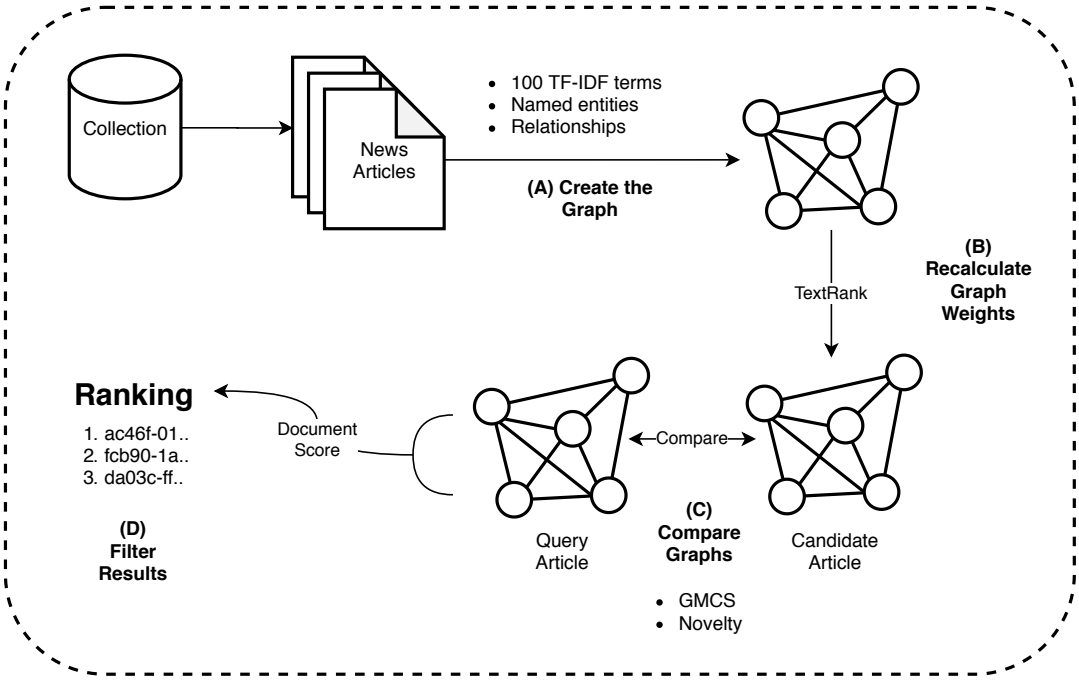

FIGURE 4.4: Overview of complete pipeline.

Chapter 5

Results

5.1 Graph Representation

The first set of questions aimed to investigate the performance of graph representations for news articles in the background linking task. An overview of effectiveness scores for different graph configurations is presented in table 5.1 and 5.2.

Model	number of terms	default	+ term position
Baseline	100	0.5217	∼
Graph	100	0.4868	**0.5326**
Graph	50	0.5020	0.5185

TABLE 5.1: NDCG@5 scores graph vs. baseline (nodes only).

The update of node weights based on the position of the node's term increased the effectiveness score for graphs using both 100 and 50 nodes. The graph that made use of 100 terms outperformed the graph that made use of 50 terms, but only when adding a score for term positions to the node's weight.

Model	# terms	default	+ paragraph	+ embedding	+ combined
Baseline	100	0.5217	∼	∼	∼
Graph	100	0.5326	0.5362	**0.5381**	0.5348
Graph	50	0.5185	0.5167	0.5089	0.5152

TABLE 5.2: NDCG@5 scores graph vs. baseline (edges).

The addition of edges to the best performing graphs from table 5.1 (here shown as defaults) positively influenced the effectiveness score when using 100 nodes. There was a negative effect for the graph that used 50 nodes. The edges based on word embeddings scored higher than the edges based on paragraph succession when using the 100-node graph, vice versa for the 50-node graph. Combining both edge weight choices resulted in a worse performance than using the individual best performing edge choice for both models.

The best performing graph configuration made use of the additional term position weights for the graph's nodes and used word embeddings for the creation of edges. An NDCG@5 score of 0.5381 was obtained and is shown in bold. Unfortunately, no significant difference between the introduced method and the baseline was found. Comparing the baseline using a two-sided t-test resulted in a p-value of

0.5759 for the best performing graph model, other configurations also did not show statistically significant differences.

5.2 TextRank

The purpose of experiment 2 was to see if the use of a 'voting' algorithm to recalculate the node weights would influence the background linking performance. Figure 5.1 compares the results per topic article for the best performing graph method with a duplicate that applied the TextRank algorithm.

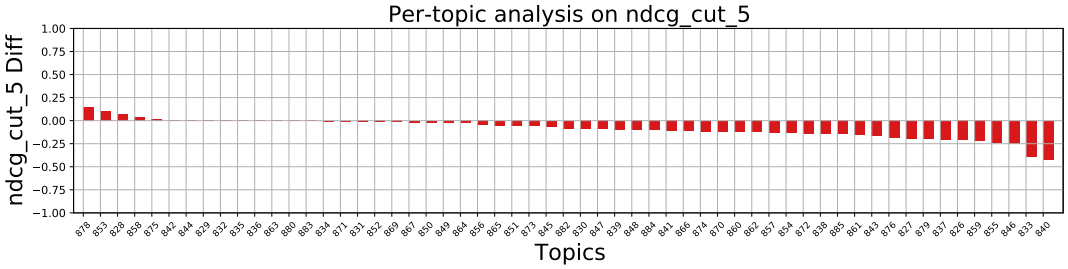

FIGURE 5.1: Difference in NDCG@5 for graph using TextRank.

No increase in effectiveness was found when using the TextRank algorithm. Moreover, the model using TextRank obtained a significantly worse score than the model without (0.4531 vs 0.5381). The model using TextRank also performed significantly worse than the baseline (p=.0220).

5.3 Named Entities

In order to investigate the influence of named entities in the graph representations, we combined the terms in our default graph method with the named entities present in a specific news article. Table 5.3 provides NDCG@5 scores for these runs.

Model	number of terms	default	+ named entities
Baseline	100	0.5217	∼
Graph	100	**0.5381**	0.5362
Graph	50	0.5089	0.4901
Graph	0	∼	0.3280

TABLE 5.3: Comparison of graph representation with named entities (default refers to highest scoring configuration until this point).

The effectiveness of the graph representation did not increase with the addition of named entities. Using only named entities also showed a considerable decrease in performance compared to the usage of only *tf-idf* terms, 0.3280 vs 0.5381/0.5089.

5.4 Novelty

Experiment 4 focused on the inclusion of new information. The effect of the novelty algorithm was tested against the (default) similarity-only measure of the greatest maximum common subgraph. NDCG@5 scores are shown in table 5.4.

Model	number of terms	default	+ novelty	+ novelty (incl. entities)
Baseline	100	0.5217	∼	∼
Graph	100	**0.5381**	0.5289	0.5271
Graph	50	0.5089	0.5091	0.5137
Graph	0	∼	∼	0.3216

TABLE 5.4: Comparison of graph representation using novelty function (default refers to highest scoring configuration until this point).

The addition of the novelty functionality did not affect the background linking performance considerably. A small decrease in effectiveness score was shown for the default graph representation and a slight increase was found in the performance of the graph with 50 nodes. This effect was stronger when named entities were incorporated. None of these differences were statistically significant.

5.5 Diversification

The final experiment measured the effect of a diversity filter for background recommendations. The described diversification led to the recommendation of 12 unlabeled articles (i.e. not retrieved by previous models) and reached an NDCG@5 score of 0.4498 instead of the original 0.5362.

5.6 Overview

Figure 5.2 shows the per topic performance of each individual method in comparison with the baseline. Overall the proposed methods scored lower than the baseline. Worst performance was achieved for topics 869, 839, and 836, whereas best performance was achieved for topics 871, 864, and 843. Article titles are visible in table 5.5.

#	Worst performance
1	Conservatives are more likely to believe that vaccines cause autism
2	Websites where children are prostituted are immune from prosecution. But why?
3	How amphetamine use may be affecting our waterways
#	Best performance
1	China plans a new moon probe in response to possible return there by U.S.
2	Sarah Palin's son, and the link between combat duty and veteran violence
3	Teen birthrate hits all-time low, led by 50 percent decline among Hispanics and blacks

TABLE 5.5: Title of best and worst performing topics.

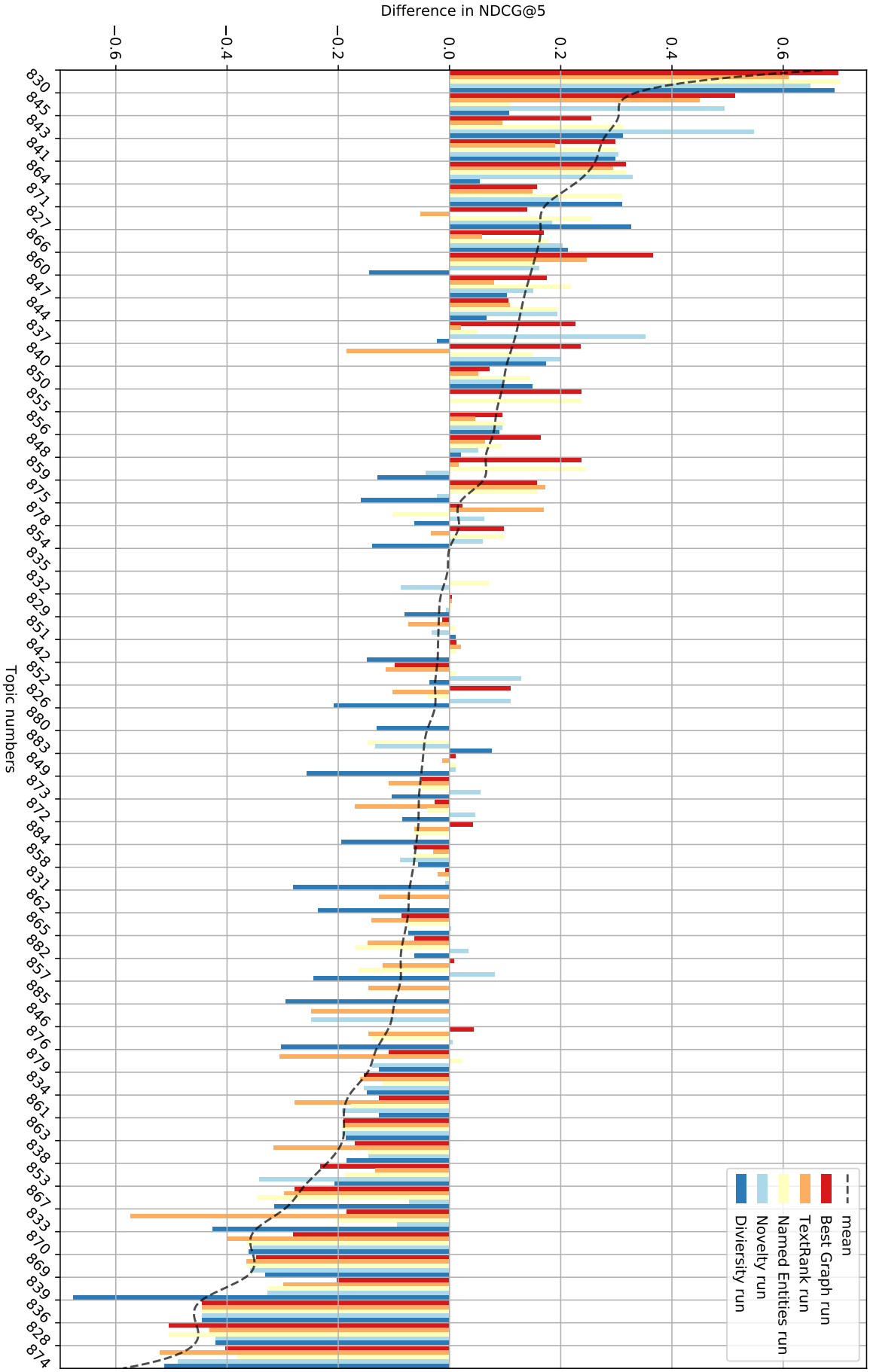

FIGURE 5.2: Per topic comparison of baseline vs introduced methods.

Chapter 6

Discussion

6.1 Interpretation of results

The objective of this study is to investigate the performance of a graph-based retrieval model for the task of background linking. Out of all tested graph configurations, the best performing graph used 100 *tf-idf* terms as nodes and derived edge weights via word embeddings. Even though this configuration obtained the highest NDCG@5 score, no significant difference was found between the graph approach and the baseline. Based on these results we have to conclude that the usage of a graph-based retrieval model does not outperform the widely used "bag-of-words" model in the task of background linking.

The foremost reason that no significant difference was obtained probably lies in the low sample size. Only 60 topic articles were available for evaluation; therefore, small differences are not powerful enough to reach statistical significance. Another reason might be the small contrast between the graph model and the baseline. Even though we tested many configurations, the one with the highest effectiveness had many features in common with the baseline and might not have been different enough to deviate considerably from the baseline recommendations. Also, to counter the increase in complexity that comes with a graph-based approach, the model relied on 100 candidate documents for the selection of background articles. These candidates were retrieved using the baseline method, and might not have provided enough opportunity for variation between predictions. Both models used the *tf-idf* score as a definition of term importance, which makes it likely that similar documents obtain a high relevance score.

The fact that the best performing graph used similar features as the baseline could indicate the superiority of these features over the alternatives we tested. All the ideas that distinguished our graph model from the baseline showed no significant improvement in effectiveness. In fact, the use of TextRank, the adoption of named entities, and the application of a diversity filter all reduced the background linking effectiveness.

6.1.1 Importance of specific nodes

The importance of graph parts was indicated by the height of the node weights. Both nodes and the connection between nodes incorporated specific weight measures. Nodes were always initialized with a variant of the *tf-idf* score and had an option to adopt an additional score based on the placement of the term in the text. This showed an increase in NDCG@5; however, this increase was not significant. It would be interesting to see how effectiveness would be affected by varying emphasis on specific weight functions. Edge weights were either based on the distance between nodes in a text or based on the distance between nodes in vector space. The highest score in effectiveness was achieved by the latter, albeit no significance was reached.

The recalculation of node weights using TextRank did not improve the NDCG of the top 5 recommended articles. A reason for this could be that the node and edge weights already were optimized before applying TextRank. In that case, the graph dynamics were not altered. Another reason could be inaccurate connection between nodes, which may be caused by the representation of connections as obtained from the word embedding. The distance words have in vector space is a mere representation of a language broad connection between the terms and may not be applicable to the relation two terms have in a specific article.

6.1.2 Novelty score

The addition of a novelty score in the relevance function did not significantly contribute to the effectiveness of the model. Neither was there a decrease in performance. This finding was not completely unexpected as the collection did not contain any duplicate documents, making it easy to retrieve documents with new information, while searching for documents that look like the topic article. Subsequently, all news articles came from the same publisher, this makes it easy to avoid documents that discuss the same exact subject in a similar way as the topic article. In order to really comprehend the contribution of the novelty score, we should test on a collection with multiple publishers and near-duplicate documents.

6.1.3 Diversity

Results of this study showed a decrease in the effectiveness of background recommendations after applying a diversity filter. It is, however, hard to obtain a clear view on performance, since a side effect of the filter is the retrieval of more unlabeled documents. In order to establish a statistically justified conclusion unlabeled documents should be annotated first. A reason for the decrease in performance could be the fact that some articles are naturally more about persons and organizations than others. Forcing articles that contain locations to higher positions might decrease the effectiveness, as documents that discuss locations are inherently irrelevant for the topic article.

6.2 Limitations and future work

The overlap between baseline and graph representation might have been a suboptimal design choice. Both methods highly depended on the *tf-idf* score of document terms for its core business. It is, therefore, harder to obtain significant differences in model performance. On the contrary, the similarity between both models also allowed to obtain a clearer view on the influence of the factors that did differ. Take the effect of different connection configurations for example.

The graph designs were limited to the use of single terms as nodes, except for the named entities (whom could consist of multiple terms - but always denoted a single concept). It would be interesting to see the performance of multiple terms per node in the form of N-grams or sentences. Another interesting direction for future research lies in the creation of graphs for individual paragraphs. News articles from established sources are often extensive, and cover a wide variety of subtopics. Background material that is relevant for one paragraph might not be for the other. Therefore, it could be easier to find relevant information for smaller pieces of text than for complete articles.

Chapter 7

Conclusion

This work investigated the effectiveness of a graph-based approach for the task of background linking. Experiments were conducted to test different graph configuration and relevance criteria. These experiments included the use of named entities, the TextRank algorithm, a novelty score, and a diversity filter. Code used to run these experiments and to generate runs for TREC 2020 can be found on GitHub [1]. The following answers summarize the findings in this thesis:

RQ 1: What is the performance of a graph-based retrieval model in the task of background linking?

While the graph-based retrieval model obtained a higher NDCG@5 score than the baseline model, no significant difference was measured. Therefore, the usage of a graph-based model does not outperform a "bag-of-words" model in the task of background linking.

RQ 1a: What is the best graph representation of news articles for this task?

The graph representation that obtained the highest background linking score used 100 tf-idf terms as nodes and used the distance between term vectors in word embeddings to obtain edges.

RQ 1b: How can we find the most important part of a graph that represents a news story?

The initial attribution of weights based on tf-idf score appeared to be the best indicator of importance. The recalculation of node weights using the TextRank algorithm resulted in a decrease in effectiveness.

RQ 2: Does the inclusion of a score for new information affect background linking effectiveness?

The effectiveness was affected only slightly by the addition of a novelty score in the relevance criterion. The background linking score dropped from 0.5381 to 0.5289 and did not show a significant difference.

RQ 3: Can we produce a diverse list of recommendations without decreasing effectiveness?

While the applied diversity filter produced a set of varying news articles, the effectiveness showed a significant decrease.

[1] https://github.com/PepijnBoers/background-linking

Bibliography

[1] W. Fox, *Writing the News, A Guide for Print Journalists*. Iowa State University Press, 2001.

[2] A. Mitchell, J. Gottfried, M. Barthel, and E. Shearer. (Jul. 7, 2016). The modern news consumer, [Online]. Available: https://www.journalism.org/2016/07/07/the-modern-news-consumer/ (visited on 04/08/2020).

[3] E. Shearer and K. E. Matsa. (Sep. 10, 2018). News use across social media platforms 2018, [Online]. Available: https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/ (visited on 04/08/2020).

[4] I. Soboroff, S. Huang, and D. Harman, "TREC 2019 news track overview", in *Proceedings of The Twenty-Eigth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA*, E. M. Voorhees and A. Ellis, Eds., ser. NIST Special Publication, vol. 1250, National Institute of Standards and Technology (NIST), 2019.

[5] P. Borlund, *The concept of relevance in IR*, 2003.

[6] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic, "Searching the web: The public and their queries", *Journal of the American Society for Information Science and Technology*, vol. 52, no. 3, pp. 226–234, 2001.

[7] M. Gupta and M. Bendersky, "Information retrieval with verbose queries", ACM, 2015.

[8] P. Yang and J. Lin, "Anserini at TREC 2018: Centre, common core, and news tracks", in *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA*, E. M. Voorhees and A. Ellis, Eds., ser. NIST Special Publication, vol. 500-331, National Institute of Standards and Technology (NIST), 2018.

[9] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond", *Foundations and Trends in Information Retrieval*, 2009.

[10] Y. Ding, X. Lian, H. Zhou, Z. Liu, H. Ding, and Z. Hou, "ICTNET at TREC 2019 news track", in *Proceedings of The Twenty-Eigth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA*, E. M. Voorhees and A. Ellis, Eds., ser. NIST Special Publication, vol. 1250, National Institute of Standards and Technology (NIST), 2019.

[11] A. Bimantara, M. Blau, K. Engelhardt, J. Gerwert, T. Gottschalk, P. Lukosz, S. Piri, N. S. Shaft, and K. Berberich, "Htw saar @ TREC 2018 news track", in *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA*, E. M. Voorhees and A. Ellis, Eds., ser. NIST Special Publication, vol. 500-331, National Institute of Standards and Technology (NIST), 2018.

[12] C. Kamphuis, F. Hasibi, A. P. de Vries, and T. Crijns, "Radboud University at TREC 2019", in *Proceedings of The Twenty-Eigth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA*, E. M. Voorhees and A. Ellis, Eds., ser. NIST Special Publication, vol. 1250, National Institute of Standards and Technology (NIST), 2019.

[13] S. Missaoui, A. MacFarlane, S. Makri, and M. Gutierrez-Lopez, "DMINR at TREC News Track", in *Proceedings of The Twenty-Eigth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA*, E. M. Voorhees and A. Ellis, Eds., ser. NIST Special Publication, vol. 1250, National Institute of Standards and Technology (NIST), 2019.

[14] K. Lu and H. Fang, "Leveraging entities in background document retrieval for news articless", in *Proceedings of The Twenty-Eigth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA*, E. M. Voorhees and A. Ellis, Eds., ser. NIST Special Publication, vol. 1250, National Institute of Standards and Technology (NIST), 2019.

[15] M. Essam and T. Elsayed, "bigIR at TREC 2019: Graph-based analysis for news background linking", in *Proceedings of The Twenty-Eigth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA*, E. M. Voorhees and A. Ellis, Eds., ser. NIST Special Publication, vol. 1250, National Institute of Standards and Technology (NIST), 2019.

[16] F. Rousseau and M. Vazirgiannis, "Main core retention on graph-of-words for single-document keyword extraction", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[17] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text", in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 404–411.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.", Stanford InfoLab, Tech. Rep., 1999.

[19] C. Florescu and C. Caragea, "PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents", in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1105–1115.

[20] A. Bougouin, F. Boudin, and B. Daille, "TopicRank: Graph-based topic ranking for keyphrase extraction", in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan: Asian Federation of Natural Language Processing, Oct. 2013, pp. 543–551.

[21] X. Wan and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge", ser. AAAI'08, Chicago, Illinois: AAAI Press, 2008, 855–860.

[22] F. Boudin, "Unsupervised keyphrase extraction with multipartite graphs", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 667–672.

[23] J. Foley, A. Montoly, and M. Peña, "Smith at TREC2019: Learning to rank background articles with poetry categories and keyphrase extraction", in *Proceedings of The Twenty-Eigth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA*, E. M. Voorhees and A. Ellis, Eds., ser. NIST Special Publication, vol. 1250, National Institute of Standards and Technology (NIST), 2019.

[24] Q.-D. Truong, T. Dkaki, J. Mothe, and P.-J. Charrel, "Gvc: A graph-based information retrieval mode.", in *Proceedings of the COnférence en Recherche d'Infomations et Applications - CORIA 2008, 5th French Information Retrieval Conference, Trégastel, France*, Jan. 2008, pp. 337–351.

[25] Z. Zhang, L. Wang, X. Xie, and H. Pan, "A Graph Based Document Retrieval Method", *Proceedings of the 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design, CSCWD 2018*, pp. 660–665, 2018.

[26] H. Bunke and K. Shearer, "A graph distance metric based on the maximal common subgraph", *Pattern Recognition Letters*, vol. 19, no. 3, pp. 255 –259, 1998.

[27] F. Hamborg, C. Breitinger, and B. Gipp, "Giveme5w1h: A universal system for extracting main events from news articles", in *Proceedings of the 13th ACM Conference on Recommender Systems, 7th International Workshop on News Recommendation and Analytics (INRA 2019)*, Copenhagen, Denmark: ACM, 2019.

[28] E. Gerritse, F. Hasibi, and A. De Vries, "Graph-embedding empowered entity retrieval", in *European Conference on Information Retrieval*, ser. ECIR '20, Springer, 2020.

[29] J. M. van Hulst, F. Hasibi, K. Dercksen, K. Balog, and A. P. de Vries, "REL: An entity linker standing on the shoulders of giants", in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20, ACM, 2020.

[30] E. M. Voorhees and A. Ellis, Eds., *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*, vol. 1250, NIST Special Publication, National Institute of Standards and Technology (NIST), 2019.

[31] E. M. Voorhees and A. Ellis, Eds., *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA*, vol. 500-331, NIST Special Publication, National Institute of Standards and Technology (NIST), 2018.